OMNIPAPER – ARQUITECTURA DE METADADOS E SUA IMPLEMENTAÇÃO NO RDF GATEWAY

Teresa Susana Mendes Pereira, Ana Alice Baptista

Universidade do Minho
Campus de Azurém, 4800-058, Guimarães, Portugal
{tpereira, analice}@dsi.uminho.pt

RESUMO

O projecto Omnipaper (Smart Access to European Newspapers) é um projecto do programa IST (Information Society Technologies) da comissão europeia, que visa fomentar um acesso comum a diferentes tipos de fontes de informação distribuída. Pretende-se chegar a um protótipo de um sistema que permita aos utilizadores (quer ocasionais, quer profissionais) um acesso estruturado, personalizado e multilíngua a todo o manancial de artigos de notícias disponibilizados pelas empresas parceiras (a que chamamos arquivos locais). Um dos aspectos fundamentais do projecto é toda a camada de metadados do sistema. Na verdade, falamos de duas camadas de metadados: (1) uma primeira camada (Local Knowledge Layer) que é adicionada aos arquivos locais e tem como principal objectivo proporcionar uma descrição semântica de todos os artigos existentes; e (2) uma segunda camada (Overall Knowledge Layer), a um nível de abstracção mais elevado que recorre á primeira para proporcionar um ambiente integrado e estruturado de navegação, e pesquisa possibilitando, quando possível uma ligação a um ambiente multilingua. Neste artigo focalizamos a nossa atenção no Local Knowledge Layer e apresentamos as estruturas de metadados utilizadas, bem como as descrições tipo RDF/XML dos diferentes géneros de artigos de notícias incluídos no protótipo. A sua implementação foi realizada, utilizando um ambiente RDF ainda em de fase de testes – o RDF Gateway. Neste momento, esta ferramenta é o suporte ao desenvolvimento do protótipo, na manipulação e tratamento de informação estruturada em RDF. Os primeiros resultados e as análises decorrentes do desenvolvimento deste protótipo, são apresentados como conclusão deste artigo.

Keywords: Metadados, RDF, Recuperação de Informação (Information Retrieval) **,** Web, Recursos, Triplos.

1. INTRODUÇÃO

Nas últimas décadas o crescimento da informação digital foi exponencial. O mesmo se verifica no crescimento da Internet. A informação está cada vez mais disponível em formato electrónico e a sua acessibilidade na Internet tem vindo a aumentar rapidamente (Paepen 2002). Este crescimento e disponibilidade contribui para a necessidade de agrupar a informação, uma vez que tanto o seu acesso como a comparação com outras fontes que se encontram geograficamente dispersas é fisicamente suportado pela Internet, contribuindo para a necessidade de integrar a informação a nível semântico.

Um dos importantes desafios dos utilizadores da Web, na pesquisa e navegação na Web reside na necessidade de organizar o imensurável número de páginas Web que surgem todos os dias a todas as horas na Internet.

É neste contexto, que surge o projecto Omnipaper, que pretende investigar formas de promover o acesso a diferentes tipos de fontes de informação distribuída. Permitindo aos utilizadores um acesso estruturado, personalizado e multilíngua a todo o conjunto de artigos de noticias disponibilizados pelas empresas parceiras (a que denominamos por arquivos locais). O Omnipaper não é um projecto de digitalização de notícias mas sim portador de notícias digitais provenientes de diversos locais e de diversos formatos.

O projecto prevê a realização em paralelo de dois protótipos seguindo duas abordagens distintas aos metadados, tanto no *Local Knowlege Layer* como no *Overall Knowlege Layer*: a abordagem *Resource Description Framework* (RDF) e a abordagem *Topic Maps* (TM).

Pretende-se com este artigo explicar, todo o trabalho desenvolvido na definição da estrutura dos metadados utilizados na descrição dos diferentes géneros de artigos que fazem parte do protótipo, assim como a abordagem RDF utilizada na sua definição.

A estrutura do artigo será composta pelas seguintes secções: na Secção 2 será feita uma breve abordagem à arquitectura do projecto Omnipaper, na Secção 3 faremos uma abordagem à tecnologia RDF contemplando o Modelo RDF e a sua Sintaxe, na Secção 4 apresentaremos estrutura de metadados definida na descrição dos artigos em RDF/XML, acompanhado de um exemplo de uma descrição do tipo RDF/XML, e das regras de descrição utilizadas, na Secção 5, será apresentado o processo de implementação e tratamento das descrições, utilizando um sistema de gestão de bases de dados nativas RDF – o RDF *Gateway*, por fim apresentaremos as conclusões e trabalho futuro na Secção 6.

2. ARQUITECTURA GERAL DO OMNIPAPER

O projecto Omnipaper teve início em Janeiro de 2002 e é composto por sete *work packages* (WPs). Cada *work package* é responsável pelo desenvolvimento de uma área muito específica do projecto encontrando-se ligadas e eventualmente dependentes.

Na arquitectura geral definida no projecto, como é ilustrado na figura 1, procedeu-se à distinção entre as camadas *Local Knowlege e Overall Knowlege Layer.*

As três *work packages* ilustradas na figura 1 (WP2, WP3 e WP5) pretendem alcançar individualmente um objectivo tecnológico distinto. No primeiro nível da camada local é analisado e testado processos de recuperar informação proveniente de fontes distribuídas (WP2). Após a combinação de um conjunto de informação cuja sua origem provém de diferentes arquivos, o seu acesso tem de ser possível e de forma uniforme (WP3).

A camada *Overall Knowlege* vai recorrer à camada anterior para proporcionar um ambiente integrado e estruturado de navegação e pesquisa, possibilitando quando possível uma ligação a um ambiente multilíngua.

No topo da camada *Overall Knowlege* é definida a interface com o utilizador que irá proporcionar uma apresentação amigável e interactiva com o utilizador de toda a camada de conhecimento (WP5).

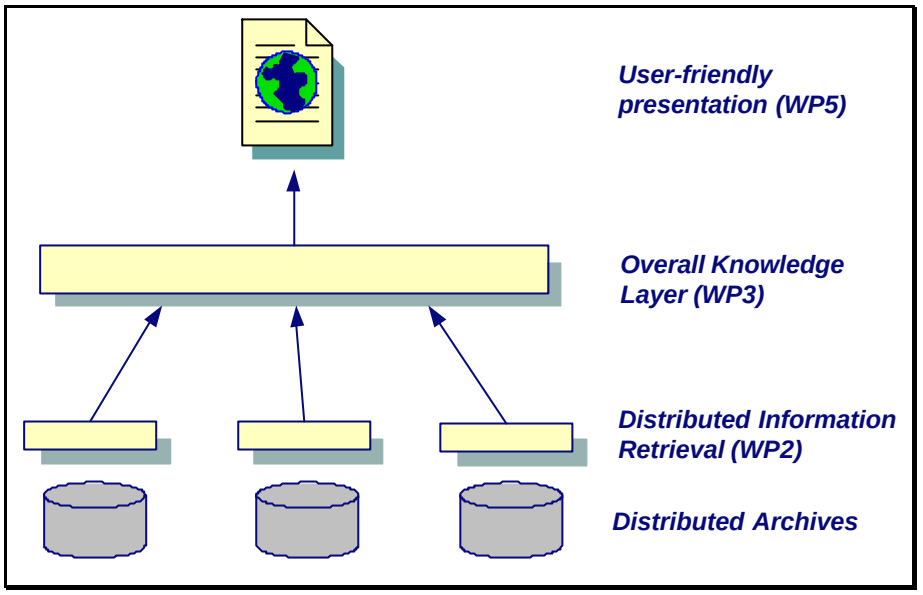

Figura 1 - Arquitectura geral do Sistema.

2.1 *Direct Retrieval Approach*

Nesta abordagem será mantida a estrutura dos dados, dos metadados, e os métodos de recuperação de informação provenientes dos arquivos locais. O que significa que não serão adicionados novos metadados aos arquivos locais.

2.2 *Local Knowlege Layer*

A implementação da camada *Local Knowlege Layer* é realizada na WP2, donde os seus resultados vão constituir o *input* para o desenvolvimento da camada *Overall Knowlege Layer* (WP3). Em ambos os níveis proceder-se-à à análise e comparação de novas tecnologias.

Novas tecnologias foram analisadas e desenvolvidas para a recuperação de informação distribuída. Pretende-se, aceder de forma estruturada e única aos arquivos locais de informação distribuída. Nesta camada é definida uma estrutura de metadados para cada artigo de notícias usando a estruturação de metadados em RDF (*Resource Description Framework*) e a tecnologia de administração do conhecimento *Topic Maps.*

No projecto estão a ser implementados em paralelo dois protótipos, seguindo duas abordagens distintas aos metadados, tanto na camada *Local* como na camada *Overall Knowledge Layer*: a abordagem *Resource Description Framework* (RDF) e a abordagem *Topic Maps* (TM). O teste cruzado dos protótipos vai contribuir para a análise e obtenção de conclusões acerca destas duas tecnologias envolvidas, em ambos os níveis.

Neste artigo aprofundamos particularmente o trabalho desenvolvido na implementação do protótipo em RDF.

2.3 *Overall Knowlege Layer*

A camada *Overall Knowlege Layer* faz a ligação entre as propriedades da informação distribuída integrada com as capacidades de definição semântica do conteúdo correspondente.

O resultado das queries executadas na camada local vão constituir o *input* para a pesquisa na camada *Overall Knowlege*. Novas técnicas vão ser analisadas e comparadas, nos dois níveis.

Técnicas que envolvem metadados são avaliadas para actuar como uma plataforma de gestão do conhecimento de forma a promover a pesquisa sobre os conteúdos assim como na sua recuperação de informação multilíngua.

A multilíngua vai ser tratado através da extracção de palavras-chave existentes e através dos metadados provenientes dos arquivos de informação heterogéneos e associa-los a um domínio específico de um *thesauri.*

A camada *Overall knowlege* contém uma rede de *thesauri* que permite aos arquivos de notícias inteligentes procurar noutros arquivos, de outros países e em outras línguas, artigos de notícias correspondentes. Contribuindo aos jornalistas e investigadores pesquisar material sobre um tópico especifico num ambiente multilingua, obtendo resultados com elevada qualidade e conteúdos relevantes.

3. TECNOLOGIA RDF (*RESOURCE DESCRIPTION FRAMEWORK)*

O *Resource Description Framework* (W3C 2003) contém, antes de tudo, um modelo para expressar semântica. Uma asserção RDF faz declarações sobre recursos, usando uma propriedade e tendo como resultado da aplicação dessa propriedade ao recurso, um valor. Uma asserção pode ser vista como um triplo composto por três elementos: propriedade (predicado), recurso (sujeito) e valor (objecto). Um recurso pode ser qualquer coisa identificável por um URI.

O modelo RDF é simplesmente um modelo de triplos, o que o torna muito poderoso, mas difícil de implementar. Por definição, a descrição usando os triplos, usando o grafo ou usando a sintaxe RDF/XML é equivalente. O *parser* RDF/XML é responsável por ler, verificar a sintaxe RDF/XML, e transformar o código escrito na sintaxe RDF/XML num conjunto de triplos e, eventualmente, num grafo RDF.

O RDF está dividido em duas partes, contendo duas especificações distintas:

(1) A *RDF Model and Syntax Specification* (RDFMSS) (W3C 2003) que é uma recomendação do W3C (W3C 2003) que contém um modelo para representar metadados RDF, bem como uma sintaxe para codificar e transportar metadados de forma a maximizar a interoperabilidade de servidores e clientes Web desenvolvidos independentemente;

(2) A *RDF Schema Specification* (W3C 2003) é uma especificação de esquemas. Com o Esquema RDF podem-se desenhar e implementar de uma forma consistente, vocabulários de metadados específicos. Estes podem ainda ser desenvolvidos no seio de outros projectos gerando, assim uma rede de esquemas de metadados. Por exemplo, determinados termos de um vocabulário a ser desenhado podem perfeitamente ser definidos como refinamentos de elementos de DC ou de outro qualquer vocabulário anteriormente definido. Na definição de metadados a utilizar na descrição dos nossos documentos foram utilizados como referência outros vocabulários, como por exemplo o NITF (*News Industry Text Format*) (IPTC 2002) e NewsML (*News Agency Implementation Guidelines*) (IPTC 2002), para além dos elementos do DC (DCMI 2003). Neste momento o *Dublin Core* é já utilizado em diversos locais do mundo e é o principal alicerce na construção e desenvolvimento do RDF.

4. ESTRUTURA DE METADADOS DEFINIDA NA DESCRIÇÃO DOS ARTIGOS EM RDF/XML

A codificação RDF/XML não é simples. Se o modelo RDF é em si mesmo de uma simplicidade extraordinária e se o rigor que permite conferir às descrições é de uma lógica matemática pura, o mesmo já não se passa com a sintaxe RDF/XML para sua implementação. De facto, num curto espaço de três anos, já vários documentos que explicam ou tentam normalizar a descrição de metadados em RDF/XML foram criados e tornados obsoletos. A DCMI já lançou vários *drafts* sobre o assunto, mas até à data em que escrevemos este artigo

ainda não há nenhuma recomendação aceite pelo DCUB. O próprio grupo de trabalho *RDF CORE* (W3C 2003) do W3C está a refazer uma nova especificação para o RDF com dois objectivos fundamentais: (1) tornar a especificação mais fácil de ler e entender e (2) eliminar alguns erros assumidos da sintaxe RDF/XML.

Neste sentido, nas nossas descrições é usado tanto quanto possível as recomendações feitas no documento *Expressing Qualified Dublin Core in* RDF/XML (Kokkelink 2002), apesar de esta ainda ser uma recomendação candidata.

De seguida apresentaremos algumas das regras aplicadas à descrição dos documentos em RDF/XML, que integram o conjunto de documentos do protótipo.

(1) **Recursos como valor de uma propriedade**
Existem duas alternativas principais para codificar um recurso como valor de uma propriedade: ou como conteúdo do elemento XML respeitante à propriedade, ou como conteúdo do atributo *rdf:Resource* associado ao mesmo elemento XML.

Por uma questão de simplicidade e clareza da descrição, codificamos os recursos que são valores de propriedades no atributo *rdf:Resource* do elemento XML correspondente à propriedade em causa.

(2) **Codificação de nodos anónimos**
Existem propriedades que estão relacionadas com uma outra, o que implica a utilização de nodos anónimos. Por exemplo se pretendermos descrever o autor do artigo, não apenas utilizando o seu nome, mas também outros dados, vamos ter que utilizar também outras propriedades que guardam o valor relativo a cada um desses tipos de dados. Assim, utilizamos primeiro a propriedade *dc:creator* que aponta para um nodo anónimo de onde vão sair novos arcos (que correspondem a propriedades) para nodos que já possuem valores. No exemplo apresentado em anexo, temos as propriedades *vCard:n*, *vCard:family*, *vCard:EMAIL* e *vCard:ORG*, que contribuem significativamente para uma melhor caracterização dos agentes, neste caso particular do autor do artigo.

A utilização do valor *"Resource"* no atributo *rdf:parseType*, indica-nos, que o conteúdo do elemento tem que ser tratado como se fosse o conteúdo de um elemento *Description*, contribuindo desta forma para uma melhor clareza e simplicidade do documento.

(3) **Utilização de *Bags* vs repetição de elementos**
Na especificação do modelo e sintaxe do RDF, um *Bag* é definido como uma lista não ordenada de recursos ou literais. *Kokkelink e Schwänzl* preferem defini-lo como uma lista cuja ordem não tem significado. Por outro lado a especificação não é clara quanto à utilização de *Bags* em contraposição à repetição de propriedades. No entanto *Kokkelink e Schwänzl* fazem uma contribuição adicional para esta distinção. Para estes autores, sempre que há uma situação de "E" lógico (*AND*), deve ser utilizada a repetição de propriedades. Alternativamente, o *Bag* deverá ser utilizado sempre que se pretende referir os valores como um conjunto, como um todo. Ou seja, o valor resultante da aplicação da propriedade é o conjunto em si mesmo. Apesar desta classificação, continua a existir uma vasta zona de fronteira e até de sobreposição, pelo que os critérios individuais do autor dos documentos RDF continuam a ter um peso significativo.

No exemplo apresentado, o elemento *subject* é codificado como um *Bag*. Foi tomada está opção, na medida em que parece evidente a utilização deste *container* uma vez que um artigo pode conter vários assuntos e portanto é o conjunto que deve ser referido.

O conjunto de todos os elementos utilizados na implementação da estrutura dos metadados em RDF/XML, pode ser consultada com mais detalhe na definição do perfil de aplicação (Pereira 2003).

De forma a ilustrar as regras descritas acima, é apresentado um exemplo aplicado à descrição de um documento em RDF/XML, que integra o conjunto de documentos do protótipo. Seguirá em anexo* a tabela de triplos e o grafo correspondente ao exemplo apresentado.

4.1 Documento RDF/XML

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/#" xmlns:dcterms ="http://purl.org/dc/terms/#"
xmlns:omn ="http://xxx/2002/06/01/schema#" xmlns:vCard="http://www.w3.org/2001/vcard-rdf/3.0#">
    <!--Description about: identificacao do recurso que esta a ser descrito -->
    <rdf:Description rdf:about="http://ultimahora.publico.pt/shownews.asp?id=170952">
        <!--TITLE-->
        <dc:title>Acidentes de trabalho mortais decrescem 21,6 por cento</dc:title>
        <dcterms:alternative>Relatório da Inspecção-Geral do Trabalho</dcterms:alternative>
        <!--CREATOR-->
        <dc:creator rdf:parseType="Resource">
            <vCard:n>Carlos</vCard:n >
            <vCard:family>Freitas</vCard:family>
            <vCard:EMAIL>cfreitas@dsi.uminho.pt</vCard:EMAIL>
            <vCard:ORG>Lusa</vCard:ORG>
        </dc:creator>
        <!--DESCRIPTION-->
        <dc:description>Os acidentes de trabalho mortais diminuíram 20,6 por cento nos primeiros seis meses
deste ano, em relação ao primeiro semestre de 2001, avança um relatório da Inspecção-Geral do Trabalho (IGT).
        </dc:description>
        <!--SUBJECT -->
        <dc:subject>
            <rdf:Bag>
                <rdf:li>Politica</rdf:li>
                <rdf:li>Trabalho</rdf:li>
            </rdf:Bag>
        </dc:subject>
        <!--PUBLISHER-->
        <!-- O atributo parserType indica-nos que o que esta a seguir deve ser tratado como codigo RDF-->
        <dc:publisher rdf:parseType="Resource">
            <vCard:ORG>Journal o Público</vCard:ORG>
            <vCard:EMAIL>publico.pt</vCard:EMAIL>
            <vCard:URL rdf:resource="http:/www.publico.pt"/>
        </dc:publisher>
        <!--DATE OF CREATION OF THE ARTICLE -->
        <dcterms:created>
            <dcterms:W3CDTF>
                <rdf:value> 2002-08-21T12:16:00 </rdf:value>
            </dcterms:W3CDTF>
        </dcterms:created>
```

```
        <!--LANGUAGE-->
        <dc:language>pt</dc:language>
        <!--FORMAT -->
        <dcterms:extent>50 KB</dcterms:extent>
        <dcterms:medium>
            <dcterms:IMT>
                <rdf:value>text/html </rdf:value>
            </dcterms:IMT>
        </dcterms:medium>
        <nitf:key-list>
            <rdf:Seq>
                <rdf:li rdf:parseType="Resource">
                    <rdf:value> nome_keyword</rdf:value>
                    <omni:key-weight>peso_keyword</omni:key-weight>
                </rdf:li>
            </rdf:Seq>
        </nitf:key-list>
    </rdf:Description>
</rdf:RDF>
```

5. IMPLEMENTAÇÃO E MANIPUÇÃO DO CÓDIGO RDF/XML

O protótipo para manipulação da camada RDF do *Local Knowledge Layer* foi realizado, utilizando o *RDF Gateway* (Chappell 2002). No entanto, antes de começarmos a utilizar esta ferramenta, o nosso primeiro contacto foi realizado com a ferramenta *Tamino XML Server* (Software AG), mas até aquela data, chegamos à conclusão que não estava preparada para trabalhar no RDF.

O *RDF Gateway* é uma ferramenta que conjuga os poderes de um servidor de HTTP com o sistema de Gestão de bases de dados nativas RDF. O conteúdo do *RDF Gateway* pode ser acedido através de um Web browser especificando o URL da aplicação que faz parte do conteúdo de uma *package* definida no *RDF Gateway*.

As aplicações são desenvolvidas numa linguagem *script* denominada *RDF Server Pages (RSP)* semelhantes às *ASP (Active Server Pages)* e as *queries* são implementadas utilizando o *RDF Query Language (RDFQL)*. Temos encontrado várias dificuldades na utilização desta ferramenta, na medida em que se trata de uma versão beta. Deste modo, estamos neste momento a analisar a viabilidade da ferramenta TAP (University 2003) como alternativa ao *RDF Gateway*.

6. CONCLUSÕES E TRABALHO FUTURO

O protótipo desenvolvido será incluído num protótipo mais vasto que cobre toda a arquitectura do *Local Knowlege Layer*. O teste cruzado e a comparação realizada ás duas novas estruturas de metadados (implementadas em Topic Maps e RDF), para serem aplicadas ás abordagens *Direct Retrivel* e *Local Knowlege* Layer vão ter um papel preponderante na definição da melhor solução para o desenvolvimento do protótipo final. Este posteriormente será integrado com o *Overall Knowlege Layer* (ver figura 1) de modo a constituírem o núcleo do sistema Omnipaper.

Numa primeira fase do protótipo, procedeu-se à descrição dos artigos de notícias digitais seguida da implementação no *RDF Gateway* dos processos necessários à sua manipulação.

No RDF Gateway, através da definição de um *datasource connection* para as nossas descrições RDF/XML, são extraídos os triplos (Sujeito, Predicado, Objecto), donde estes serão posteriormente armazenados numa base de dados nativa.

Pretende-se com este protótipo pesquisar e navegar sobre os metadados definidos na descrição dos artigos de notícias e analisar a eficiência do RDF na recuperação de informação proveniente de diferentes fontes de informação distribuída.

Na próxima fase proceder-se-á implementação do módulo de transformação, ou seja, este módulo consiste no mapeamento dos elementos de metadados existentes nos arquivos locais para a estrutura dos elementos de metadados definidos no sistema Omnipaper. Nesta primeira fase, aos arquivos locais não lhes é imposta uma definição da estrutura dos elementos de metadados, ou seja a implementação deste módulo passou primeiro pela análise da estrutura dos elementos de metadados de cada arquivo local assim como os esquemas utilizados, seguido do mapeamento dos elementos de metadados mais relevantes para o sistema Omnipaper.

De forma a responder ás necessidades de informação do utilizador, será feita calculada a relevância dos artigos de notícias, permitindo retornar ao utilizador os artigos mais relevantes segundo a pesquisa efectuada, assim como analisar o comportamento do utilizador durante a pesquisa de forma melhorar as características de navegação.

REFERÊNCIAS

Chappell, G. R., Derrish; Michal , Aaron (2002). Intellidimension Gateway.

DCMI (2003). Dublin Core Metadata Initiative.

IPTC (2002). “NewsML - News Agency Implementation Guidelines.” .

IPTC (2002). NITF - News Industry Text Format.

Kokkelink, S. S., Roland (2002). “Expressing Qualified Duclin Core in RDF/XML.” .

Paepen, B. E., Jan; Schranz, Markus; Tscheligi, Manfred (2002). Omnipaper: Bringing Electronic News Publishing to a next Level Using XML and Artificial Inteligence. ELPUB2002, Karlovy Vary, Czech Republic.

Pereira, T. B., Ana Alice, Yaginuma, Tomoko (2003). Perfil de Aplicação Implementado no Projecto Omnipaper. III Congresso Luso-Moçambicano de Engenharia, Maputo, Moçambique.

Software AG, T. X. C. Tamino XML Server.

University, S. (2003). TAP.

W3C (2003). RDF.

W3C (2003). RDF Schema.

W3C (2003). RDF Sintax and Grammar.

W3C (2003). RDFCore Working Group. **2003**.

W3C (2003). World Wide Web Consortium.